How Viral Are Viral Videos?

Christian Bauckhage
Fraunhofer IAIS
Bonn, Germany

Fabian Hadiji and **Kristian Kersting**
TU-Dortmund University
Dortmund, Germany

Abstract

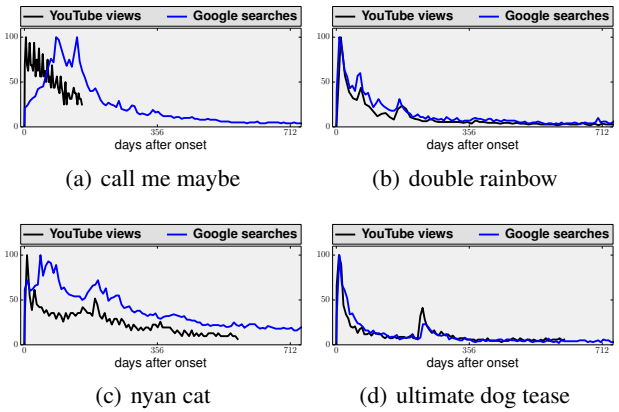
Within only a few years after the launch of video sharing platforms, viral videos have become a pervasive Internet phenomenon. Yet, notwithstanding growing scholarly interest, the suitability of the viral metaphor seems not to have been studied so far. In this paper, we therefore investigate the attention dynamics of viral videos from the point of view of mathematical epidemiology. We introduce a novel probabilistic model of the progression of infective diseases and use it to analyze time series of YouTube view counts and Google searches. Our results on a data set of almost 800 videos show that their attention dynamics are indeed well accounted for by our epidemic model. In particular, we find that the vast majority of videos considered in this study show very high infection rates.

Introduction

Viral videos have become a staple of the social Web. The term refers to videos that are uploaded to video sharing sites such as YouTube, Vimeo, or Blip.tv and more or less quickly gain the attention of millions of people.

Viral videos mainly contain humorous content such as bloopers in television shows (e.g. *boom goes the dynamite*) or quirky Web productions (e.g. *nyan cat*). Others show extraordinary events caught on video (e.g. *battle at Kruger*) or contain political messages (e.g. *kony 2012*). The arguably most prominent example, however, is the music video *Gangnam style* by PSY which, as of January 2015, has been viewed over 2 billion times on YouTube. Yet, while the recent surge in viral videos has been attributed to the availability of affordable digital cameras and video sharing sites (Grossman 2006), viral Web videos predate modern social media. An example is the *dancing baby* which appeared in 1996 and was mainly shared via email.

The fact that videos became Internet phenomena already before the first video sharing sites appeared suggests that collective attention to viral videos may spread in form of a contact process. Put differently, it seems reasonable to surmise that attention to viral videos spreads through the Web very much as viruses spread through the world. Indeed, the times series shown in Fig. 1 support this intuition. They show exemplary developments of YouTube view counts and



(a) call me maybe (b) double rainbow

(c) nyan cat (d) ultimate dog tease

Figure 1: Outbreak data related to 4 different viral videos.

Google searches related to recent viral videos and closely resemble the progress of infection counts often observed in epidemic outbreaks. However, although viral videos attract growing research efforts, the suitability of the viral metaphor was apparently not studied systematically yet. In this paper, we therefore ask to what extend the dynamics in Fig. 1 can be explained in terms of the dynamics of epidemics?

This question extends existing viral video research which, so far, can be distinguished into two broad categories: On the one hand, researchers especially in the humanities and in marketing, ask for what it is that draws attention to viral videos (Burgess 2008; Southgate, Westoby, and Page 2010). In a recent study, Shifman (2012) looked at attributes common to viral videos and, based on a corpus of 30 prominent examples, identified six predominant features, namely: focus on ordinary people, flawed masculinity, humor, simplicity, repetitiveness, and whimsical content. However, while he argues that these attributes mark a video as incomplete or flawed and therefore invoke further attention or creative dialogue, the presence of these key signifiers does not imply virality. After all, there are millions of videos that show these attributes but never attract significant viewership

Another popular line of research, especially among data scientists, therefore consists in analyzing viewing patterns of viral videos. For instance, Figueiredo et al. (2011) found that the temporal dynamics of view counts of YouTube videos

seem to depend on whether or not the material is copyrighted. While copyrighted videos (typically music videos) were observed to reach peak popularity early in their lifetime, other viral videos had been available for quite some time before they experienced sudden significant bursts in popularity. In addition, the authors observed that these bursts depended on external factors such as being listed on the YouTube front page. The importance of external effects for the viral success of a video was also noted by Broxton et al. (2013) who found that viewership patterns of YouTube videos strongly depend on referrals from sites such as Facebook or Twitter. In particular, they observed that 'social' videos with many outside referrals rise to and fall from peak popularity much quicker than 'less social' ones.

Sudden bursts in view counts seem to be suitable predictors of a video's future popularity (Crane and Sornette 2008; Pinto, Almeida, and Goncalves 2013; Jiang et al. 2014). In fact, it appears that initial view count statistics combined with additional information as to, say, video related sharing activities in other social media, allow for predicting whether or not a video will 'go viral' soon (Shamma et al. 2011; Jain, Manweiler, and Choudhury 2014). Yet, Broxton et al. (2013) point out that not all 'social' videos go viral and not all viral videos are indeed 'social'.

Given this interest in video related time series analysis, it is surprising that the viral metaphor has not been scrutinized from this angle. To the best of our knowledge, the most closely related work is found in a recent report by Cintro-Arias (2014) who attempted to match an intricate infectious disease model to view count data for the video *Gangnam style*. We, too, investigate the attention dynamics of viral videos from the point of view of mathematical epidemiology and present results based on a data set of more than 800 time series. Our contributions are of theoretical and empirical nature, namely:

1) we introduce a simple yet expressive probabilistic model of the dynamics of epidemics; in contrast to traditional approaches, our model admits a closed form expression for the evolution of infected counts and we show that it amounts to the convolution of two geometric distributions

2) we introduce a time continuous characterization of this result; major advantages of this continuous model are that it is analytically tractable and allows for the use of highly robust maximum likelihood techniques in model fitting as well as for easily interpretable results

3) we fit our model to YouTube view count data and Google Trends time series which reflect collective attention to prominent viral videos and find it to fit well.

Our work therefore constitutes a data scientific approach towards viral video research. However, it is model- rather than data driven. This way, we follow arguments brought forth, for instance, by Bauckhage et al. (2013) or Lazer et al. (2014) who criticized the lack of interpretability and the 'big data hubris' of purely data driven approaches for their potential of over-fitting and misleading results.

Our presentation proceeds as follows: Next, we review concepts from mathematical epidemiology, briefly discuss approaches based on systems of differential equations, and introduce the probabilistic model that forms the basis for our study; mathematical details behind this model are deferred to the Appendix. Then, we present the data we analyzed and discuss our empirical results. We conclude by summarizing our approach, results, and implications of our findings.

Modeling Viral Process

Mathematical models of epidemic processes play a crucial role in many disciplines. In medicine, they help studying the population dynamics of infectious diseases (Britton 2010; Lloyd and May 2001); economists use them to trace and predict the diffusion of innovations or marketing messages (Dover, Goldberg, and Shapira 2012; Leskovec, Adamic, and Huberman 2007); and, in the wake of social media, increasing efforts are spent on modeling information cascades in Web-based social networks (Adar and Adamic 2005; Bauckhage 2011; Budak, Agrawal, and Abbadi 2010; Leskovec, Adamic, and Huberman 2007; Leskovec, Backstrom, and Kleinberg 2009; Yang and Leskovec 2011).

Each of these instances addresses a surprisingly pervasive phenomenon: An agent (a virus, a rumor, an urge to buy a product, etc.) spreads in form of a contact process and thus cascades through a network of interlinked entities (people, computers, blogs, etc.). At the onset of the agent's activity, many, if not most, of the entities are *susceptible* to its effects but only a few are actually *infected*. As time progresses, susceptible entities that are in contact with infected ones may themselves become infected. Infected entities may either remain infected, *recover*, become susceptible again, or even be removed from the population.

Crucial characteristics of such an epidemic are its infection rate, its recovery rate, or the number of newly infected entities per unit of time. They help assessing the progression or final outbreak size of the process and can inform contagion or dissemination strategies (Barthelemy et al. 2004; Newman 2002). Similarities in the spread of diseases and rumors have been noted for long (Dietz 1967) and led to several applications of epidemic modeling in the context of Web technologies. Examples include attempts to predict the diffusion of messages in bulletin boards (Kubo et al. 2007) or approaches towards forecasting collective interest in Web-based services or content (Bauckhage 2011; Cannarella and Spechler 2014; Ribeiro 2014).

Epidemic dynamics within large populations are often modeled using *compartment models*. These assume the population to be divided into disjoint fractions of those who are susceptible (S) to an epidemic, those who are infected (I), and those who have recovered (R). Some models consider additional compartments but, in any case, assume an individual to belong to one group only. Transitions between groups are constrained by the structure of the model; for instance, SIR models are concerned with transitions of the form $S \rightarrow I \rightarrow R$.

Differential Equations

Traditionally, the dynamics of an SIR epidemic are characterized in terms of systems of coupled, non-linear differential equations (Britton 2010). While such approaches have been used in social media analysis before (Bauckhage 2011;

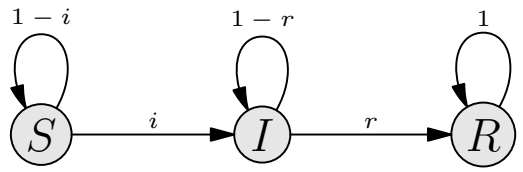
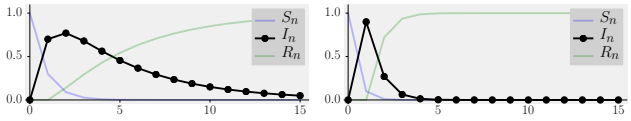


Figure 2: Markov chain model of an SIR epidemic.



(a) $i = 0.7, r = 0.2$ (b) $i = 0.9, r = 0.8$

Figure 3: Examples of the temporal evolution of fractions of susceptible, infected, and recovered individuals (S_n, I_n, and R_n) according to the Markov chain in Fig. 2.

Cannarella and Spechler 2014; Ribeiro 2014; Cintron-Arias 2014), we note that they are known not to allow for closed form solutions of the temporal behavior of any of their constituent parts. In other words, they do not provide a simple expression for the function $I(t)$ which describes the temporal progression of infected counts in an epidemic outbreak. This poses certain technical difficulties in fitting such a model to observed data and may necessitate elaborate data cleaning (Cannarella and Spechler 2014). In order to avoid these kind of difficulties, our work in this paper is based on a probabilistic model which we introduce below.

Markov Processes

The compartmentalized nature of SIR models also suggests to describe their dynamics in terms of Markov processes. Indeed, Markov processes are another popular approach in mathematical epidemiology where they are usually applied to represent the behavior of individuals (Britton 2010). Here, however, we focus on the population level.

We consider the homogeneous, time discrete Markov chain in Fig. 2 where the infection rate $0 < i < 1$ and the recovery rate $0 < r < 1$ indicate transition probabilities between the compartments or states of the model. As for any discrete Markov chain, the temporal dynamics can be expressed in terms of a recursive matrix-vector equation which, in our case, amounts to

$$\begin{bmatrix} S_n \\ I_n \\ R_n \end{bmatrix} = \begin{bmatrix} 1-i & 0 & 0 \\ i & 1-r & 0 \\ 0 & r & 1 \end{bmatrix} \begin{bmatrix} S_{n-1} \\ I_{n-1} \\ R_{n-1} \end{bmatrix} \tag{1}$$

where we define the initial state distribution vector as

$$\begin{bmatrix} S_0 \\ I_0 \\ R_0 \end{bmatrix} = \begin{bmatrix} 1-\epsilon \\ \epsilon \\ 0 \end{bmatrix} \tag{2}$$

for some $0 \leq \epsilon < 1$.

Figure 3 shows examples for how this model behaves over 15 iterations if $\epsilon = 0$. Apparently, this conceptually simple Markov process is able to explain a wide range of infection dynamics (from stretched to peaky). However, it is a time discrete model and a time continuous model would be more tractable analytically. Our empirical analysis in this paper is therefore based on the following

Theorem 1. *Given the time discrete homogenous Markov model in* (1)*, the temporal distribution of percentages of infected individuals I_n can be characterized in terms of a continuous probability density function $f(t)$ which is the convolution of two exponential distributions, i.e.*

$$f(t) = \int_0^t \lambda\, e^{-\lambda\tau}\, \alpha\, e^{-\alpha(t-\tau)}\, d\tau. \tag{3}$$

The rate parameters of the two exponentials are given by

$$\lambda = -\ln(1-i) \tag{4}$$

and

$$\alpha = -\ln(1-r), \tag{5}$$

respectively. The convolution in (3) *has a closed form solution which amounts to*

$$f(t) = \begin{cases} \frac{\alpha}{\alpha-\lambda}\, \lambda\, e^{-\lambda t} + \frac{\lambda}{\lambda-\alpha}\, \alpha\, e^{-\alpha t} & \text{if } \lambda \neq \alpha \\ \lambda^2\, t\, e^{-\lambda t} & \text{if } \lambda = \alpha. \end{cases} \tag{6}$$

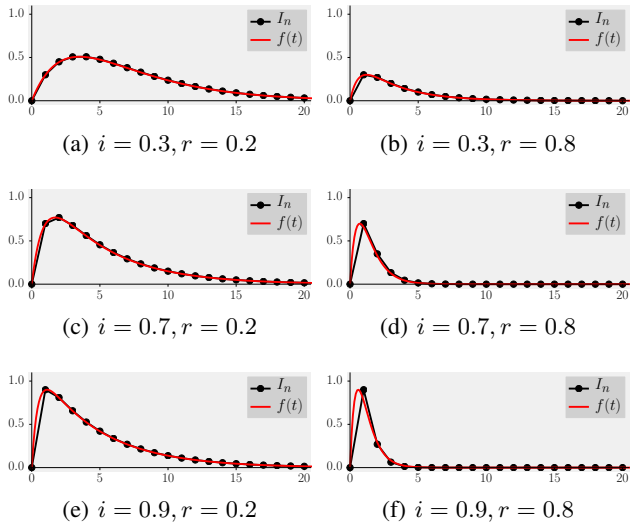
To the best of our knowledge, the result in (6) provides a novel characterization of the temporal dynamics of infection counts in epidemic outbreaks and has not been studied in this context before.

As the derivation of this result is rather technical, we defer it to the Appendix. Here, we point out the following favorable properties of our model: since $f(t)$ in (6) is a probability density function, the model lends itself to statistical methods and Bayesian reasoning. Moreover, it immediately allows for the use of maximum likelihood approaches for parameter estimation based on empirical data. In contrast to other probability density function such as the Weibull or the LogNormal distribution which have previously been used to characterize collective attention processes (Bauckhage, Kersting, and Hadiji 2013), the abstract shape parameters λ and α in (6) are in a one-to-one relation with the infection rate i and the recovery rate r of an infectious process and therefore admit an intuitive and physically plausible interpretation of analysis results.

Below, we fit the model in (6) to YouTube view count data and Google Trends time series which reflect attention dynamics to viral videos. First, however, we introduce the data sets considered.

Data Sets

The empirical basis for our study of the viral dynamics of viral videos consists of two data sets which we describe in the following.



(a) $i = 0.3, r = 0.2$ (b) $i = 0.3, r = 0.8$

(c) $i = 0.7, r = 0.2$ (d) $i = 0.7, r = 0.8$

(e) $i = 0.9, r = 0.2$ (f) $i = 0.9, r = 0.8$

Figure 4: Examples of how the continuous density $f(t)$ in (6) fits discrete time series I_n of fractions of infected individuals generated by the Markov chain in Fig. 2.

Table 1: Statistics for the 10 most viewed videos in the CMU data set; Δt indicates the delay (in days) between onset and peak of the corresponding view count time series.

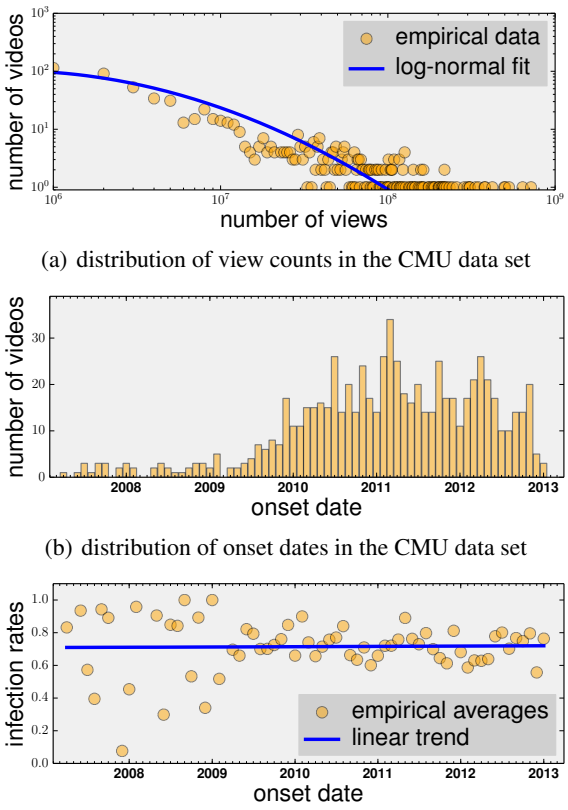
video	# views	onset	Δt
J. Bieber – baby	717189290	2010-05-03	12
J. Lopez – on the floor	638903361	2011-03-09	1
Eminem – love the way	529923543	2010-08-13	1
Lady Gaga – bad romance	501031279	2009-12-05	1
Shakira – waka waka	495809532	2010-06-23	10
LMFAO – party rock	420312033	2011-07-01	8
Eminem – not afraid	380252443	2010-06-14	1
Pitbull – rain over me	378923711	2011-07-25	6
J. Bieber – never say never	347277927	2010-06-15	10
B. Mars – lazy song	333510376	2011-05-16	8

CMU Viral Video Data

Recently, Jian et al. (2014) applied Hidden Markov Models to forecast future interest in a video based on view counts and meta information such as user comments. Their data was gathered from YouTube and constitutes the largest publicly available such data set to date. It currently consists of 2,526 records of time series of view counts, comments, and ratings. In our analysis, we restrict ourselves to the view counts.

The most popular item in this data set is Justin Bieber's music video *baby* which had amassed more than 700 million views by the time the data was gathered in early 2013. At the same time, the least viewed videos in the collection set had been viewed less than 100 times. As it seems unreasonable to consider videos of such low view counts to be viral, we decided to ignore videos of less than 500,000 views. This left us with with time series of view counts for 726 different YouTube videos.

Table 1 lists basic statistics as to the top ten most viewed videos in this subset and we note that each of these videos is



(a) distribution of view counts in the CMU data set

(b) distribution of onset dates in the CMU data set

(c) average infection rates over time in the CMU data set

Figure 5: Basic statistics for the videos in the CMU data set.

a copyrighted music video. In order to verify whether or not there are other biases in the data, we therefore performed a series of elementary tests.

Figure 5(a) shows that while the majority of videos attracted moderately many views, a few videos have indeed been watched more 100 million times. Overall, the distribution apparent from the figure resembles a power law distribution and thus hints at a preferential attachment or rich-get-richer effect with respect to the popularity of already popular videos. Upon closer examination, however, we found a LogNormal distribution to give a more accurate fit which, in turn, is not unusual in the context of social media and Web content and may indicate preferential attachment dynamics (Mitzenmacher 2004; Szabo and Huberman 2010).

Figure 5(b) shows a histogram of onset times. For every month in the observation period (from July 2007 to March 2013), it counts the number of videos first viewed in this month. Apparently, most of the videos in this set were uploaded between 2010 and 2012 which may be due to the fact that Jian et al. (2014) selected many videos based on the fact that they had been mentioned on Ray William Johnson's YouTube channel which back then regularly reviewed currently popular videos.

Finally, in a slight look-ahead to our main analysis below, Fig. 5(c) plots average infection rates over time. The two most prominent features apparent from this plot are that

Table 2: Statistics for the videos in the Google Trends data set; Δt indicates the delay (in weeks) between onset and peak of the corresponding search frequency time series.

video	# views	onset	Δt
PSY – gangnam style	2192408766	2012-08-26	5
charlie bit me	807708021	2007-12-23	16
C.R. Jepsen – call me	626950559	2012-03-11	13
evolution of dance	288994817	2006-05-14	4
ultimate dog tease	164867115	2011-05-01	2
mysterious ticking noise	157572495	2007-04-15	15
david after dentist	127449363	2009-02-01	2
nyan cat	116009795	2011-04-17	7
talking twin babies	111482944	2011-03-27	1
how animals eat food	110705037	2013-04-07	2
kony 2012	100075009	2012-03-04	1
chocolate rain	99805423	2007-07-15	4
ghost elevator prank	93687638	2012-11-25	1
jk wedding	87908181	2009-07-19	2
lonelygirl15	77665196	2006-09-03	2
harlem shake	76396739	2013-02-10	2
surprised kitty	74766489	2009-11-29	1
laughing baby	74557681	2011-02-27	1
numa numa	55736078	2005-02-13	8
leave britney alone	48917728	2007-09-09	1
i like turtles	46888725	2010-08-22	4
ninja cat	45224785	2008-09-14	3
dramatic chipmunk	42626245	2007-06-17	3
leeroy jenkins	41808467	2005-05-15	28
cinamon challenge	41718813	2012-01-15	3
double rainbow	40801688	2010-07-04	2
keyboard cat	38161315	2009-05-03	3
scarlet takes tumble	28848978	2008-10-19	7
chuck norris split	23455740	2013-12-15	1
charlie the unicorn	23312150	2006-09-03	16
omg cat	22854405	2010-03-21	2
otters holding hands	20204375	2007-03-25	2
kittens inspired by kittens	18689110	2009-02-08	2
share it maybe	18569317	2012-07-08	1
diet coke mentos	17388675	2006-06-04	2
will it blend	17388604	2007-07-08	1
the last lecture	17124865	2008-04-06	1
united breaks guitars	14455669	2009-07-05	1
christian the lion	11948840	2008-07-20	2
boom goes the dynamite	9013576	2005-06-05	2
autotune the news	8629101	2010-08-01	1
standing cat	6887942	2010-04-04	2
overly attached girlfriend	2916780	2012-06-03	3
here it goes again	n/a	2006-08-06	4
lazy sunday	n/a	2005-12-18	2
pale kid raps fast	n/a	2011-01-16	2
XXX	n/a	2007-10-21	7

average infection rates are rather high and hardly change over time; the slope of the fitted linear trend is only ever so slightly positive (0.00016). All in all, even though the most viewed videos in this data set exclusively consist of professionally produced and copyrighted music videos, it therefore appears that this data set provides a representative sample of viral videos that may be used for further analysis.

Google Trends Data

In addition to the direct engagement data (view counts) in the CMU data set, we also consider search logs related to 50 different viral videos which we obtained from Google Trends in late 2014. This service provides statistics about queries submitted to Google's search engine and is increasingly used as a proxy in research on attention dynamics (see, e.g. Bauckhage et al. (2013) or Cannarella and Spechler (2014) and the discussions therein). Among others, it supplies weekly summaries of how frequently a query has been used in a specific country. This allowed us to gather search data for the nine largest English speaking countries in order to be able to spot possible regional differences in the adaption of viral videos. The countries we considered are: Australia, Canada, Great Britain, Ireland, Nigeria, New Zealand, Singapore, the USA, and South Africa.

We note that, in contrast to the time series in the CMU data set, time series obtained from Google Trends are normalized such that the peak search activity for a query corresponds to a value of 100. Data obtained from Google Trends therefore only indicates relative search frequencies rather than absolute interest in a viral video. In the time series plots shown throughout this paper we normalized the CMU data correspondingly for better comparability.

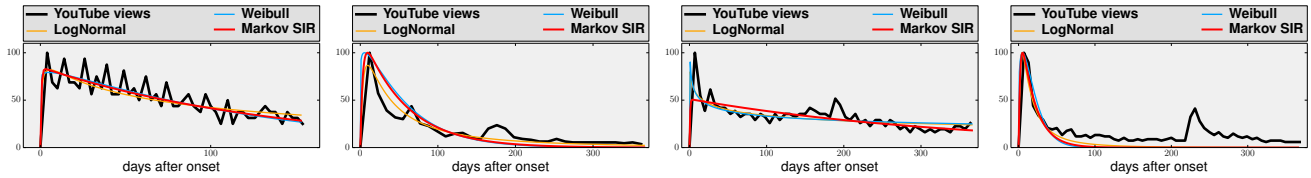
Table 2 shows basic statistics for the videos contained in this set. The view counts were determined from the corresponding YouTube pages and reflect the situation as of January 2015; we note that four videos for which we retrieved Google Trends data had been removed from YouTube and that we garbled the name of one of these (XXX) because of its controversial content. The onset dates and times to peak popularity in the table were determined from the Google Trends time series (averaged over all countries). In agreement with previous analyses of YouTube view count data (Crane and Sornette 2008; Pinto, Almeida, and Goncalves 2013; Jiang et al. 2014) we observe that for Google search data, too, most videos listed in the table reach peak popularity early in their lifetimes; that is, 25% peaked in the first week and 54% peaked within two weeks.

Empirical Analysis

In our empirical analysis, we applied maximum likelihood methods to fit the *Markov* SIR distribution introduced in (6) to the time series described above. For baseline comparison, we also fitted *LogNormal* and *Weibull* distributions which have been found to represent the dynamics of collective attention processes (Bauckhage, Kersting, and Hadiji 2013; Bauckhage, Kersting, and Rastegarpanah 2014; Szabo and Huberman 2010).

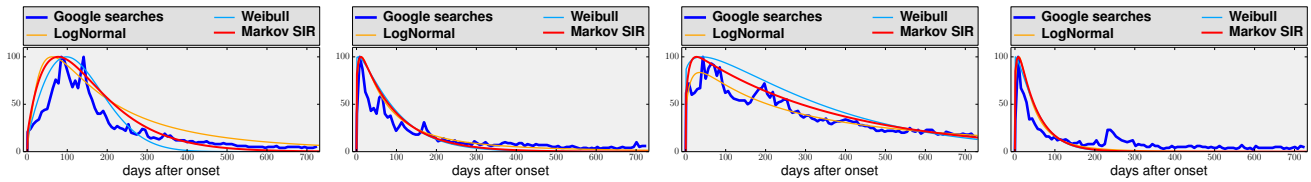
Model fitting was done with respect to the raw data without any pre-processing such as temporal smoothing. For each time series, we applied CUSUM statistics (Page 1954) to determine onset times and fitted to the data from the onset onward. If there were several onsets, such as in the case of the *ultimate dog tease video* (see Fig. 1), we fitted models to each of the corresponding sub-sequences.

Figures 6 and 7 present qualitative examples of fits to YouTube view count and Google search data, respectively.



(a) call me maybe (b) double rainbow (c) nyan cat (d) ultimate do tease

Figure 6: Examples of YouTube view count time series and fitted models (for the *ultimate dog tease* video, we only show fits to the first onset to avoid visual clutter). Each of the three models accounts reasonably well for the general trends of growing and declining attention apparent from the time series. Yet, only the Markov SIR model introduced in this paper allows for an interpretation of these dynamics in terms of an epidemic process.



(a) call me maybe (b) double rainbow (c) nyan cat (d) ultimate do tease

Figure 7: Examples of Google search frequency time series and fitted models. Each of the three models accounts reasonably well for the general trends apparent in the time series.

We note that for both data sets the time series fluctuate about a notable general trend of growth and decline. In accordance with the theoretical considerations in this paper as well with results reported in the previous literature (Bauckhage, Kersting, and Hadiji 2013), we observe that each of the three models provides reasonably accurate descriptions of global trends. While the Weibull and the LogNormal model explain these trends in terms of rather abstract concepts, the Markov SIR model introduced in this paper allows for an intuitive interpretation in terms of an epidemic process.

To quantify goodness-of-fit of the three models, we computed Effron's measure $0 \leq R^2 \leq 1$ where $R^2 = 1$ would indicate a perfect fit. In case of the YouTube view count data, we found average values of 0.57, 0.45, and 0.48 for the Markov SIR, the Weibull, and the LogNormal model, respectively. For the Google search frequency data, we found averages of 0.65, 0.62, and 0.60.

These quantitative results confirm the qualitative results in Figs. 6 and 7 in that they indicate that all three models are well capable of characterizing general trends in the data. Interestingly, the epidemic model introduced in this paper provides the overall best fits. This suggests that attention dynamics to viral videos may in indeed be explained in terms of a viral process.

Having fitted the Markov SIR model to the data, i.e. having determined the most likely shape parameters λ and α of the density in (6), allowed us to use equations (4) and (5) to solve for the infection and recovery rates of the assumed infection processes. Figure 8 plots infection rates versus recovery rates determined from the YouTube view count data in the CMU data set. It is noticeable that attention to most videos in this data set evolves with high infection rates and low recovery rates (see the histograms on top and to the right of the main panel). Since high infection rates translate to rapid initial growth of viral cascades, our findings provide an explanation for previous reports by Figueiredo et al. (2011) and Jiang et al. (2014) who observed that viral videos tend to reach peak popularity early in their lifetime.

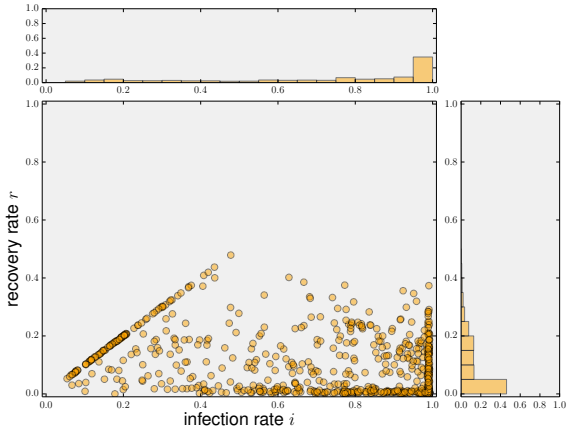


Figure 8: Scatter plot of infection rates i versus recovery rates r determined form YouTube view count time series. For the majority of the 726 videos, we observe high infection rates and low recovery rates (42% have an infection rate higher than 0.9).

Figure 9 shows similar results determined from the data set of video related Google searches. Again, we observe that attention to most videos in this data set evolves with high infection rates and low recovery rates. However, on average, infection rates are slightly smaller and recovery rates are

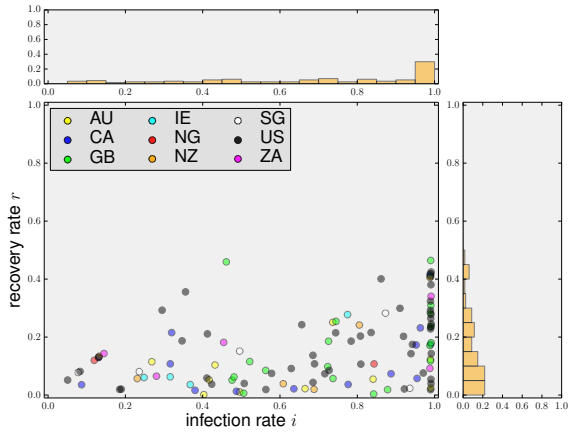


Figure 9: Scatter plot of infection rates i versus recovery rates r determined from country specific time series of Google searches for viral videos. For most of the corresponding attention processes, we observe high infection rates and low recovery rates (35% have an infection rate higher than 0.9) yet there seem to be no country specific viral dynamics.

Table 3: Most viral videos in the CMU data set.

video	Shifman (2012) category
united breaks guitars	ordinary people, humor, *music*
Eminem - love the way	*professional music video*
splendiferous barfing cup	whimsical
dramatic chipmunk	whimsical
Ken Block's Gymkhana 3	repetitiveness, *cars*
Japan earthquake	*news event*
Jerrod Niemann - lover, lover	*professional music video*
Usher - there goes my baby	*professional music video*
pacman frog	whimsical
Nicki Minaj - super bass	*professional music video*

Table 4: Most viral videos in the Google Trends data set.

video	Shifman (2012) category
pale kid raps fast	ordinary people, *music*
kony 2012	*political*
laughing baby	whimsical , humor
leave britney alone	ordinary people, whimsical
ghost elevator prank	humor
chuck norris split	whimsical, humor
united breaks guitars	ordinary people, humor, *music*
share it maybe	*music*
surprised kitty	whimsical, *cats*
the last lecture	ordinary people

higher than in the case of the CMU data. This agrees with the qualitative examples in Fig. 7 which indicate that attention to viral videos expressed in Google search data peaks later than attention apparent from YouTube view counts. This is an interesting effect that merits further studies which, for now, we leave to future work. We also note that the figure does not reveal any country specific viral dynamics. This is to say that the data points shown in the figure do not form any obvious, statistical significant clusters that would indicate specific video related search behaviors for any of the countries considered here.

Finally, after estimating infection rates for each of the videos considered here, we determined the top ten videos of highest viral pressure, i.e. with the highest infection rates, and attempted to categorize their content according to Shifman's typology (Shifman 2012). Tables 3 and 4 list our results for the YouTube view count and Google search data, respectively. Apparently, only one video (*united breaks guitars*) appears in both lists. We also observe that music videos, either by amateurs or professionals, feature prominently. Finally, it appears as if whimsical content, too, is well represented among particularly viral viral videos.

Conclusion

In this paper, we were concerned with the question "How viral are viral videos?". We approached this problem from the point of view of time series analysis and considered a Markov model of the population dynamics of an SIR epidemic. In a theoretical contribution, we showed that the temporal behavior of infection counts due to this model can be characterized using a continuous probability density that results from convolving two exponential distributions. In contrast to traditional epidemic models, our approach therefore allows for a closed form solution, is thus analytically tractable, and lends itself to maximum likelihood approaches in model fitting.

We applied this model to analyze almost 800 time series which reflect pattern of growing and declining collective attention to individual viral videos and found it to account well for the general dynamics present in this data. In fact, our approach gave better fits than baseline models from the previous literature. Our empirical findings therefore suggest that it is indeed reasonable to assume that attention to viral videos spreads in form of a viral process.

We also showed that, in contrast to previous baseline models, the shape parameters of the probability density we derived are directly related to the infection and recovery rates of the underlying epidemic process. This allowed us to determine infection and recovery rates from YouTube view count and Google search data and we found that most videos in our data sets were highly infectious. Among the top ten most infectious videos, musical content abounds but whimsical content, too, seems to be well represented.

Given the methods and results presented in this paper, there are several directions for future work. First and foremost, we foresee applications in trend prediction. Mathematical epidemiology is a mature field and concepts such as the basic reproduction number allow for estimating if, say, a viral outbreak is imminent or what the final outbreak size will be. Corresponding techniques can be used to predict the future attention a video will receive and could thus inform the planning of marketing campaigns. Of course, infection and recovery rates would to be estimated reliably from only a few initial observations as to view counts or search engine queries. But here, too, our approach offers new possibilities.

As our model is probabilistic, it lends itself to Bayesian techniques and reasoning under uncertainty.

Using Bayesian techniques for model fitting and trend prediction, one may, for instance, impose a Beta-prior on the parameters i and r of the epidemic model and update the current best estimate accordingly once new information arrives. Corresponding work is underway and we hope to report results soon.

Appendix: Proof of Theorem 1

In this appendix, we discuss the mathematical derivation of the results in equations (4), (5), and (6) that were presented above.

In addition to the recursive relation $\boldsymbol{\pi}_n = \boldsymbol{P}\,\boldsymbol{\pi}_{n-1}$ in (1), the behavior of the Markov chain in Fig. 2 can also be expressed in closed form $\boldsymbol{\pi}_n = \boldsymbol{P}^n\,\boldsymbol{\pi}_0$. Accordingly, the Markovian SIR model allows for closed form expressions for the dynamics of each of its compartments. For instance, unrolling the recursion in (1), we find that the percentage of susceptible individuals after n steps of the process is given by

$$S_n = (1-i)^n\,S_0. \tag{7}$$

Likewise, after more tedious but straightforward algebra, we find the percentage of infected individuals after n steps to amount to

$$I_n = \sum_{k=0}^{n-1}(1-r)^k\,i\,(1-i)^{n-1-k}\,S_0 + (1-r)^n\,I_0. \tag{8}$$

As this expression is slightly more involved than the one in (7), we will further simplify it. To this end, we first of all assume that the initial percentage I_0 of infected individual is zero which, according to (2) implies that the initial percentage S_0 of susceptible individuals is one. Equation (8) then simplifies to

$$I_n = \sum_{k=0}^{n-1}(1-r)^k\,i\,(1-i)^{n-1-k}. \tag{9}$$

Second of all, we recast the result in (9). To this end, we multiply the right hand side by a factor of 1 which we express as

$$1 = \frac{r\,(1-r)}{r\,(1-r)} \tag{10}$$

and obtain

$$I_n = \frac{1}{r\,(1-r)}\sum_{k=0}^{n-1} r\,(1-r)^{k+1}\,i\,(1-i)^{n-(k+1)} \tag{11}$$

$$= \frac{1}{r\,(1-r)}\sum_{j=1}^{n} r\,(1-r)^{j}\,i\,(1-i)^{n-j} \tag{12}$$

$$= \frac{1}{r\,(1-r)}\sum_{j=1}^{n} g_r[j]\,g_i[n-j]. \tag{13}$$

This expression is equivalent to the one in (9) but we can now recognize it as a scaled convolution of two geometric distributions $g_r[n]$ and $g_i[n]$.

Third of all, we provide a continuous characterization of the discrete convolution in (13). To this end, we note that discrete geometric distributions can be expressed in terms of continuous exponential distributions. For example, $g_i[n]$ in (13) can be written as

$$g_i[n] = i\,(1-i)^{n-1} = (1-e^{-\lambda})(e^{-\lambda})^{n-1} \tag{14}$$

$$= (e^{\lambda}-1)\,e^{-\lambda n} \tag{15}$$

$$= \frac{e^{\lambda}-1}{\lambda}\,\lambda\,e^{-\lambda n} \tag{16}$$

where $\lambda = -\ln(1-i)$. Similarly, we have

$$g_r[n] = r\,(1-r)^{n-1} = \frac{e^{\alpha}-1}{\alpha}\,\alpha\,e^{-\alpha n} \tag{17}$$

where $\alpha = -\ln(1-r)$.

These results then allow for characterizing the temporal dynamics of the percentage of infected individuals in terms of a time continuous function. That is, instead of I_n, we may consider

$$I(t) = C\cdot\int_0^t \lambda\,e^{-\lambda\tau}\,\alpha\,e^{-\alpha(t-\tau)}\,d\tau \tag{18}$$

where C is a time independent scaling factor given by

$$C = \frac{1}{r\,(1-r)}\cdot\frac{e^{\lambda}-1}{\lambda}\cdot\frac{e^{\alpha}-1}{\alpha}. \tag{19}$$

In other words, up to scaling, the temporal evolution of the percentage of infected individuals in a Markovian SIR model is governed by the convolution of two exponential distributions for which we have

$$f(t) = \int_0^t \lambda\,e^{-\lambda\tau}\,\alpha\,e^{-\alpha(t-\tau)}\,d\tau \tag{20}$$

$$= \frac{\alpha}{\alpha-\lambda}\,\lambda\,e^{-\lambda t} + \frac{\lambda}{\lambda-\alpha}\,\alpha\,e^{-\alpha t}. \tag{21}$$

Finally, we note that (21) is undefined if $\lambda = \alpha$. Yet, for this case, too, there exists a solution. To show this, we recall that the Laplace transform of the convolution of two functions is given by the product of the Laplace transforms of the individual functions. Since the Laplace transform of an exponential density is

$$L(s) = \int_0^{\infty} e^{-st}\,\lambda\,e^{-\lambda t}\,dt = \frac{\lambda}{\lambda+s} \tag{22}$$

the Laplace transform of the convolution of two identical exponentials amounts to $\lambda^2/(\lambda+s)^2$. This, however, is the Laplace transform of the probability density function of a gamma distribution with a shape parameter of $k=2$. This establishes that, for $\lambda=\alpha$, the convolution in (20) produces

$$f(t) = \frac{\lambda^k}{\Gamma(k)}\,t^{k-1}\,e^{-\lambda t} = \lambda^2\,t\,e^{-\lambda t}. \tag{23}$$

Acknowledgements

This work was carried out within the research project *SoFWIReD* and funded by the Fraunhofer ICON initiative. Fabian Hadiji was supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project A6 "Resource-efficient Graph Mining".

References

Adar, E., and Adamic, A. 2005. Tracking Information Epidemics in Blogspace. In *Proc. Int. Conf. on Web Intelligence*. IEEE/WIC/ACM.

Barthelemy, M.; Barrat, A.; Pastor-Satorras, R.; and Vespignani, A. 2004. Velocity and Hierarchical Spread of Epidemic Outbreaks in Scale-free Networks. *Physical Review Letters* 92(17):178701.

Bauckhage, C.; Kersting, K.; and Hadiji, F. 2013. Mathematical Models of Fads Explain the Temporal Dynamics of Internet Memes. In *Proc. Int. Conf. on Weblogs and Social Media*. AAAI.

Bauckhage, C.; Kersting, K.; and Rastegarpanah, B. 2014. Collective Attention to Social Media Evolves According to Diffusion Models. In *Proc. Int. Conf. on WWW*. ACM.

Bauckhage, C. 2011. Insights into Internet Memes. In *Proc. Int. Conf. on Weblogs and Social Media*. AAAI.

Britton, T. 2010. Stochastic Epidemic Models: A Survey. *Mathematical Biosciences* 225(1):24–35.

Broxton, T.; Interian, Y.; Vaver, J.; and Wattenhofer, M. 2013. Catching a viral video. *J. of Intelligent Information Systems* 40(2):241–259.

Budak, C.; Agrawal, D.; and Abbadi, A. E. 2010. Limiting the Spread of Misinformation in Social Networks. In *Proc. Int. Conf. on WWW*. ACM.

Burgess, J. 2008. "All Your Chocolate Rain Are Belong to Us"? Viral Video, YouTube and the Dynamics of Participatory Culture. In Lovink, G., and Niederer, S., eds., *The Video Vortex Reader: Responses to YouTube*. Amsterdam: Institute of Network Cultures. 101–110.

Cannarella, J., and Spechler, J. 2014. Epidemiological Modeling of Online Social Network Dynamics. *arXiv : 1401.4208 [cs.SI]*.

Cintron-Arias, A. 2014. To Go Viral. *arXiv:1402.3499 [physics.soc-ph]*.

Crane, R., and Sornette, D. 2008. Viral, Quality, and Junk Videos on YouTube: Separating Content from Noise in an Information-Rich Environment. In *Proc. Spring Symp. on Social Information Processing*. AAAI.

Dietz, K. 1967. Epidemics and Rumors: A Survey. *J. of the Royal Statistical Society A* 130(4):505–528.

Dover, Y.; Goldberg, J.; and Shapira, D. 2012. Network Traces on Penetration: Uncovering Degree Distribution from Adoption Data. *Marketing Science* 31(4):689–712.

Figueiredo, F.; Benevenuto, F.; and Almeida, J. 2011. The Tube over Time: Characterizing Popularity Growth of YouTube Videos. In *Proc. Int. Conf. on Web Search and Data Mining*. ACM.

Grossman, L. 2006. How to get famous in 3500 seconds. Time Magazine.

Jain, P.; Manweiler, J.; and Choudhury, A. A. R. 2014. Scalable Social Analytics for Live Viral Event Prediction. In *Proc. Int. Conf. on Weblogs and Social Media*. AAAI.

Jiang, L.; Miao, Y.; Zhang, Y.; Lan, Z.; and Hauptmann, A. 2014. Viral Video Style: A Closer Look at Viral Videos on YouTube. In *Proc. Int. Conf. on Multimedia Retrieval*. ACM.

Kubo, M.; Naruse, K.; Sato, H.; and Matubara, T. 2007. The Possibility of an Epidemic Meme Analogy for Web Community Population Analysis. In *Proc. int. Conf. on Intelligent Data Engineering and Automated Learning*.

Lazer, D.; Kennedy, R.; King, G.; and Vesignani, A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176):1203–1205.

Leskovec, J.; Adamic, L.; and Huberman, B. 2007. The Dynamics of Viral Marketing. *ACM Tans. on the Web* 1(1):5.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the Dynamics of the News Cycle. In *Proc. Inf. Conf. on Knowledge Discovery and Data Mining*. ACM.

Lloyd, A., and May, R. 2001. How Viruses Spread Among Computers and People. *Science* 292(5520):1316–1317.

Mitzenmacher, M. 2004. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics* 1(2):226–251.

Newman, M. 2002. Spread of Epidemic Disease on Networks. *Physical Review E* 66(1):016128.

Page, E. 1954. Continuous Inspection Scheme. *Biometrika* 41(1–2):100–115.

Pinto, J.; Almeida, J.; and Goncalves, M. 2013. Using Early View Patterns to Predict the Popularity of YouTube Videos. In *Proc. Int. Conf. on Web Search and Data Mining*. ACM.

Ribeiro, B. 2014. Modeling and Predicting the Growth and Death of Membership-Based Websites. In *Proc. Int. Conf. on WWW*. ACM.

Shamma, D.; Yew, J.; Kennedy, L.; and Churchill, E. 2011. Viral Actions: Predicting Video View Counts Using Synchronous Sharing Behaviors. In *Proc. Int. Conf. on Weblogs and Social Media*. AAAI.

Shifman, L. 2012. An Anatomy of a YouTube Meme. *New Media & Society March* 14(2):187–203.

Southgate, D.; Westoby, N.; and Page, G. 2010. Creative Determinants of Viral Video Viewing. *Int. J. of Advertising* 29(3):349–368.

Szabo, G., and Huberman, B. 2010. Predicting the Popularity of Online Content. *Comm. of the ACM* 53(8):80–88.

Yang, J., and Leskovec, J. 2011. Patterns of Temporal Variation in Online Media. In *Proc. Int. Conf. on Web Search and Data Mining*. ACM.